FAB-Attack: Fabric-Aware Adversarial Attacks on Person Detectors under Motion Blur

Anonymous Author(s)
Submission Id: 1664

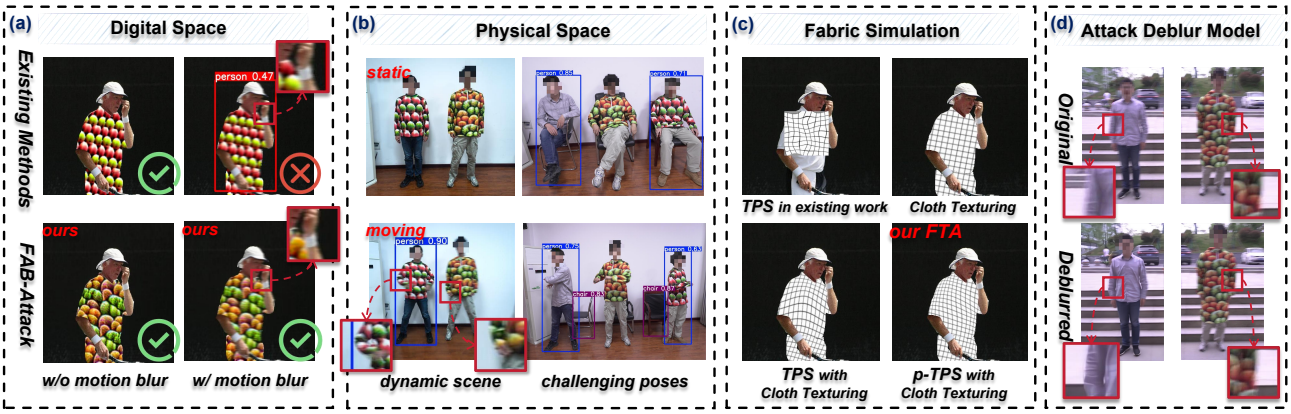


Figure 1: (a) In the digital space, existing methods degrade significantly under even mild blur. (b) In the physical space, existing methods fail under motion-induced blur (left) and clothing deformation caused by different body poses (right). (c) Existing methods simulate clothing deformation by applying random Thin Plate Spline (TPS) to the patch. In contrast, we use physics-inspired TPS on the cloth region to model realistic clothing deformation. (d) Furthermore, we explicitly account for the camera's internal deblurring model, attacking it to interfere with downstream tasks.

ABSTRACT

Physical adversarial attacks on person detectors reveal critical vulnerabilities in safety-critical vision systems such as autonomous driving and surveillance. While recent methods enhance attack efficacy and robustness, they often neglect realistic garment deformations and motion blur, limiting real-world performance. In this work, we propose **FAB-Attack** (**F**abric-**a**ware and **B**lur-resistant Attack), a new adversarial attack that simulates realistic garment deformation during training and targets both person detectors and image deblurring models. To enhance attack effectiveness under varying clothing deformations, we introduce a Fabric-aware Texture Appliance (FTA) module, which applies adversarial textures to clothing regions and simulates realistic fabric dynamics via physics-inspired TPS. To better emulate real-world conditions, we develop a differentiable pipeline incorporating motion blur and deblurring processes. Moreover, we demonstrate the stability of low-frequency information during motion blur's generation and removal. Based on this insight, we design a frequency band separation mechanism that suppresses high-frequency components in adversarial patterns to enhance further robustness against motion blur. Experimental results demonstrate that our approach achieves SOTA performance, reducing AP to 25.2% on the COCO dataset and achieving a 94.4% ASR in the real world under severe motion blur.

CCS CONCEPTS

• **Security and privacy → Usability in security and privacy**.

KEYWORDS

Adversarial Texture, Physical Attack, Motion Blur

1 INTRODUCTION

Person detection is a safety-critical computer vision task with widespread applications in surveillance systems and autonomous driving infrastructures. Given its pivotal role in safety assurance, recent years have witnessed growing research efforts on physical adversarial attacks (PAA) against person detectors [9, 12–16, 26, 28, 36, 37, 39]. These attacks typically craft adversarial patterns in the digital domain and transfer them to the physical world, where they are captured by cameras and re-digitized for detection. A key challenge is preserving attack effectiveness through this digital-physical domain transformation.

Physical carriers for deploying adversarial patterns from the digital to the physical domain have evolved from discrete patches to full-body textures. Patch-based methods[9, 12, 15, 26, 28, 36, 37, 39] print patterns and affix them to the wearer's torso, whereas texture-based methods [13, 14, 16] integrate patterns directly into clothing. The latter has gained attention for broader body coverage and improved robustness across viewpoints. Despite recent advances in adapting texture-based attacks to the physical domain—e.g., constraining printable colors via the Non-Printability Score [24], incorporating rotations and scalings during training [1, 27], and simulating cloth deformations with Thin Plate Spline (TPS) [13, 14]—existing methods still fall short of achieving high attack success rates in the physical world (see Sec. 4.3.2).

We identify three primary factors contributing to this limitation: (1) Low-fidelity training domain: Current methods optimize adversarial textures on low-quality 3D synthetic data [13] or patch-based training paradigm [14], creating a substantial domain gap to real

scenes. (2) Unrealistic fabric deformation: To model non-rigid clothing deformations, existing methods apply random Thin Plate Spline (TPS) during training [13, 14], which poorly align with real-world garment deformation trends and underperform in physical space. (3) Motion-blur blind zone: We observe that even mild motion blur severely degrades attack efficacy, yet no prior work tackles motion blur robustness. Moreover, modern cameras often incorporate image deblurring models as a preprocessing step for downstream tasks, and attacks targeting such pipelines remain unexplored.

In this paper, we propose **Fabric-Aware and Blur-Resistant Adversarial Attack (FAB-Attack)**, a novel method addressing these challenges via two key innovations. *First*, we introduce **Fabric-Aware Texture Appliance (FTA)**, a physics-grounded approach that renders fabric-realistic textures on clothing in real-world datasets, thereby bridging the gap between training and deployment domains while accurately simulating garment deformation. FTA is realized through attaching scale-specific texture onto cloth region masks. During this process, we introduce physics-inspired Thin Plate Spline (p-TPS) to facilitate realistic fabric deformation on the texture aligning with physical conditions. *Second*, to address the motion-blur blind zone and enhance the robustness of our attack, we integrate a random motion blur generator along with a deblurring model into our attack pipeline. This allows us to simulate the full sensor capture-to-image restoration process commonly employed in smart cameras. Additionally, our theoretical analysis reveals that low-frequency information remains stable during motion blur's generation and removal. Leveraging this insight, we introduce **High Frequency Loss**, significantly improving the attack's effectiveness under dynamic conditions.

We conduct experiments on the COCO dataset [20] and a collected set of 5,000 real-world captured images under diverse conditions. The results confirm that FAB-Attack demonstrates robust and superior attack performance in both global and local motion blur scenarios in real-world environments, causing the failure of both person detection and deblurring models across various smart camera setups. Notably, our approach consistently outperforms existing methods under both white-box and diverse black-box settings, achieving superior average precision (AP) and attack success rate (ASR). Our contributions are summarized as follows:

- We propose Fabric-Aware Texture Appliance (FAT), the first method that synthesizes realistic fabric textures on real-world pedestrian datasets, ensuring adversarial texture that realistically simulates the physical deformation of clothing.
- We reveal the stability of low-frequency information during the blur and deblurring processes and propose a high-frequency loss to impose physical constraints on the texture. We simultaneously attack the deblurring models to impact downstream tasks.
- We conduct comprehensive evaluations showing FAB-Attack's superior efficacy—reducing AP to 25.2% on COCO and achieving 94.4% ASR in the real world, effectively attacking detectors under motion blur while suppressing deblurring models.

2 RELATED WORK

2.1 Physical Adversarial Attack

Compared to digital adversarial attacks [4, 25, 38], physical adversarial attacks (PAAs) have garnered significant research attention due to their ability to threaten real-world DNN-based systems. Following the seminal work by Sharif et al. [24] who demonstrated successful attacks on face recognition systems through physically realizable perturbations, numerous PAA methods have emerged across various vision tasks, including classification [3], object detection [28], segmentation [22], and image captioning [41]. These methods typically generate digital perturbations that are then materialized into physical entities and deployed in real-world scenarios.

A major challenge in PAA lies in addressing the *digital-to-physical domain gap* caused by environmental variations and sensor distortions. Early approaches [1, 27] incorporated physical simulation (e.g., rotation, scaling) during training to enhance perturbation robustness. Subsequent works introduced specialized constraints, such as the Non-Printability Score (NPS) [24] for color reproducibility and BRDF-based texture modeling [13] for material-aware pattern generation. Recent work by Wei et al. [35] made a significant advancement by introducing the *physical-to-digital domain gap*, focusing on camera-specific effects, and pioneering the use of differentiable Image Signal Processing (ISP) pipelines to simulate sensor color calibrations. However, existing methods predominantly address static imaging conditions while neglecting dynamic artifacts like motion blur – a prevalent phenomenon in safety-critical scenarios that significantly impacts attack effectiveness.

2.2 PAA against Person Detection Models

In this work, we focus on physical adversarial attacks against person detectors. Existing relative methods can be broadly categorized into two classes: patch-based attacks [9, 15, 28, 36, 37], and texture-based attacks [13, 14, 16]. The former employs localized and regular-shaped patches as the adversarial medium while the latter utilizes clothing as the medium. While patch-based attacks are simple in design and use, they face the critical segment-missing problem [14] and are effective only from the very frontal viewpoint. In contrast, texture-based methods have become increasingly attractive due to their larger perturbation coverage, providing robust attack performance across various viewpoints and human poses. However, current texture-based approaches predominantly rely either on patch training paradigm [14] or low-fidelity synthetic 3D data [13, 16], resulting in a significant digital-to-physical domain gap. Moreover, the random Thin Plate Splines (TPS) used in these methods fails to accurately model the realistic deformation of clothing. To address these challenges, we propose Fabric-aware Texture Appliance (FTA), which leverages emulated physics-inspired TPS to accurately apply adversarial textures onto clothing regions in real-world datasets. This approach substantially mitigates the digital-to-physical discrepancy during training, achieving superior attack effectiveness in physical environments.

2.3 Image Deblurring

Recent deep learning-based methods, leveraging powerful convolutional neural networks (CNNs) and transformer-based end-to-end architectures, have made significant strides in image restoration tasks, particularly blind image deblurring [7, 19, 30, 33, 40]. These approaches predominantly focused on pixel-level representations for deblurring. More recent approaches have also considered the influence of the blur kernel and incorporated frequency-domain

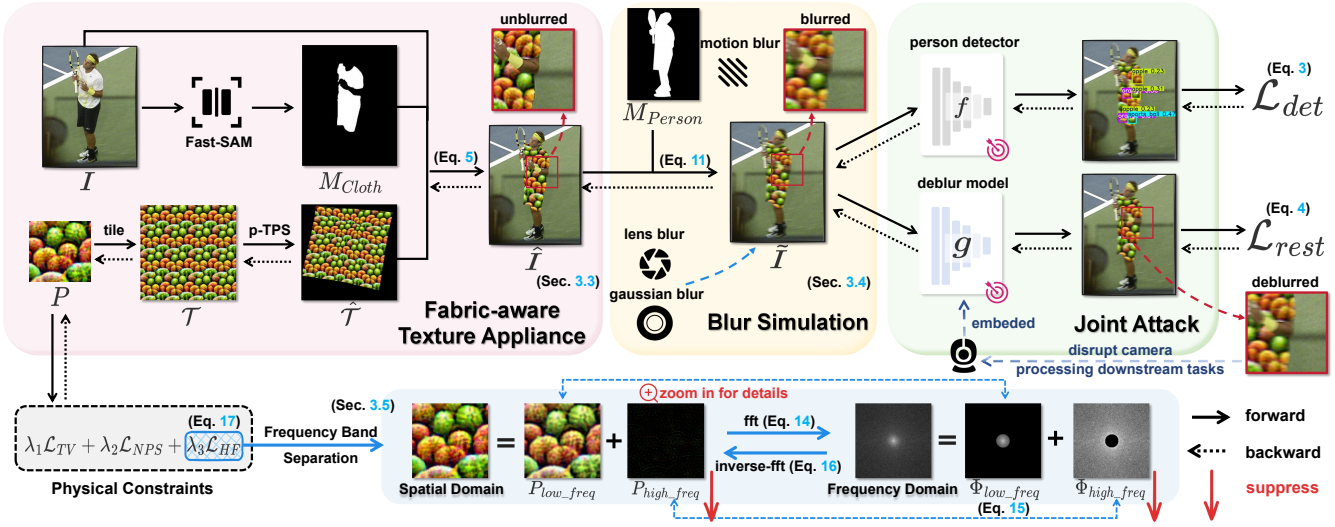


Figure 2: Overview of FAB-Attack. Given an input image I, we first obtain the clothing mask M_{Cloth} using Fast-SAM[42]. The adversarial patch p is then tiled into texture $\mathcal{T}$. The image I, mask M_{Cloth}, and texture $\mathcal{T}$ are fed into the proposed FTA to generate a textured image $\hat{I}$. This image is further processed by a blur simulation module to produce a blurred version $\tilde{I}$. The blurred image $\tilde{I}$ is fed into both the person detector and the deblurring model for a joint attack. During optimization, a High Frequency Loss ($\mathcal{L}_{HF}$) is introduced to suppress unstable high-frequency components.

information to better exploit these cues, as demonstrated by DeepRFT [21]. These deblurring models, recognized for their effectiveness in recovering fine details and high-frequency components, have been widely integrated into camera preprocessing pipelines. Additionally, they have shown dynamic applicability, significantly enhancing the performance of various downstream vision tasks, including object detection [11, 29] and recognition [6]. However, despite these advancements, the vulnerability of image restoration models to adversarial attacks remains underexplored, which poses a significant concern in safety-critical applications.

3 METHODS

3.1 Problem Formulation

Our objective is to design a physically realizable adversarial attack against person detection models and image restoration models simultaneously, while maintaining robustness across dynamic scenario with various motion blur in digital and physical space.

Our approach optimizes a base patch P, which is tiled to form an adversarial texture $\mathcal{T} \in \mathbb{R}^{\infty\times\infty\times 3}$. Given an input image I containing multiple persons, an attack sample is generated through the following process: the texture $\mathcal{T}$ is first applied to the clothing regions specified by binary masks $M_{Cloth} = \{m_i\}_{i=1}^{I}$, where $m_i \in \mathbb{R}^{H\times W}$. Subsequently, the motion blur effects are simulated to mimic degradation during image acquisition. The overall process can be formulated as:

$$\tilde{I} = B(F(I, \mathcal{T}, M_{Cloth}), M_{Person}), \tag{1}$$

where $F(\cdot)$ denotes our proposed Fabric-aware Texture Appliance (FTA), $M_{Person} = \{m_j\}_{j=1}^{J}$ represents masks for each person in the image I, $B(\cdot)$ randomly selects between localized body blur (using M_{Person} masks) or global motion blur, and $\tilde{I}$ represents the final attack sample generated by the process.

Given the attack sample $\tilde{I}$, our adversarial objective contains three key components and can be formulated as below:

$$\min_{P} \mathbb{E}_{I\sim\mathcal{D}} \left[\mathcal{L}_{det} + \mu\mathcal{L}_{rest} + \lambda\Omega\right]. \tag{2}$$

(1) **Detection Attack** $\mathcal{L}_{det}$: Minimizes the detection loss for blurred-and-restored adversarial images:

$$\mathcal{L}_{det} = \sum_{i=1}^{M} c_i(f(\tilde{I})), \tag{3}$$

where function $f(\cdot)$ represents a pre-trained person detection model, outputting detection boxes with a maximum number of M, and c_i denotes taking the confidence score of the i-th detection box.
(2) **Restoration Attack** $\mathcal{L}_{rest}$: Minimizes the discrepancy between input and output of the restoration model:

$$\mathcal{L}_{rest} = \|g(\tilde{I}) - \tilde{I}\|_2, \tag{4}$$

where function $g(\cdot)$ denotes a pre-trained image restoration model.
(3) **Physical Constraints** Ω: Combines three regularizers:

- *Total Variation Loss* $\mathcal{L}_{TV}$: Enforces spatial smoothness.
- *Non-Printability Score* $\mathcal{L}_{NPS}$: Ensures fabric printable colors.
- *High Frequency Loss* $\mathcal{L}_{HF}$: Optimize adversarial efficacy in low-frequency bands.

3.2 Overview

We proposed Fabric-Aware and Blur-Resistant Attack (FAB-Attack), an effective texture-based physical adversarial attack method (see Fig. 2) that is resilient to fabric deformation and motion blur. To align training data with real-world garment deployment and better simulate the physical deformations of real garments, we introduce Fabric-Aware Texture Appliance (FTA) to create realistic attack

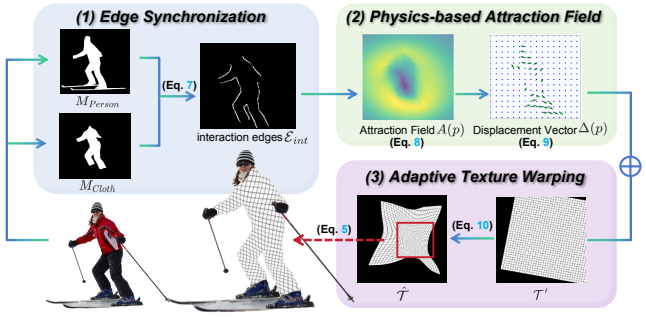


Figure 3: The pipeline of our physics-inspired TPS for simulating realistic fabric deformations.

samples. FTA begins by extending the patch into a texture, followed by scaling based on the human body size. Next, we apply random adjustments to brightness and contrast, as well as random cutouts. We mask the transformed textures with cloth segmentation labels (see *Appendix Section D* for data acquisition). During this process, we introduce a physics-inspired Thin Plate Spline (p-TPS) to ensure realistic fabric deformation, aligning the textures with physical conditions. Following this, the textured image is passed through a random dynamic blur function to simulate motion blur, producing a blurred image that reflects real-world degradation during image capture. Then the blurred images are processed by a shake-dropped person detection model [15] and image deblurring model separately.

The optimization process aims to minimize the following objectives: the confidence score of the person detection model, the restoration performance of the deblurring model, as well as the non-printability score (NPS) [24] and total variation (TV) of the patch. To further enhance robustness to both motion blur and deblurring, we introduce a frequency-domain regularization, High Frequency Loss, which suppresses the high-frequency components of the patch that are most susceptible to such degradations.

3.3 Fabric-aware Texture Appliance

Existing paradigms [14] or low-fidelity synthetic 3D simulations [13] introduce significant discrepancies between digital training settings and real-world testing conditions. Additionally, these methods often rely on random TPS for simulating clothing deformation. To overcome these limitations, we propose the Fabric-aware Texture Appliance (FTA), a method designed to realistically simulate adversarial textures on human clothing within widely-used real-world datasets. The entire process of FTA can be expressed as:

$$\hat{I} = (p\text{-TPS}\left(\mathcal{S}(\mathcal{T}, \text{bbox}(I))\right) \odot M_{Cloth}) + (I \odot (1 - M_{Cloth})), \quad (5)$$

where $\mathcal{T}$ is the base texture, $\text{bbox}(I)$ is the bounding box of the person in the image I, $\mathcal{S}()$ represents the scaling of the texture based on the bounding box, p-TPS$(\cdot)$ denotes the proposed **physics-inspired thin-plate spline**, and $\odot$ denotes element-wise multiplication. The detailed implementation of p-TPS is as follows.

To realistically simulate physical fabric deformation on clothing, we introduce a physics-inspired deformation field derived from edge attraction forces. The rationale behind this edge attraction mechanism stems from two fundamental observations regarding real-world fabric deformation: (1) *Non-uniform curvature distribution*: bending and wrinkling predominantly occur near garment boundaries, where physical constraints (e.g., seams and body contact points) lead to higher stress concentrations. (2) *Nonlinear force propagation*: the magnitude of deformation decays inversely proportional to the square of the distance from boundary anchors, as evidenced in classical fabric draping models [2]. The proposed p-TPS operates through three progressive stages:

(1) **Edge Synchronization.** To begin, we first detect coherent edge regions between the person's body mask and the clothing mask using a dual-stream Scharr operator. The mathematical formulation for edge detection is as follows:

$$\mathcal{E}_p = \sigma(|\mathcal{F}_{\text{Scharr}}(M_{Person})|), \quad \mathcal{E}_c = \sigma(|\mathcal{F}_{\text{Scharr}}(M_{Cloth})|), \quad (6)$$

where $\mathcal{F}_{\text{Scharr}}$ represents the Scharr convolution operation. The magnitude of the resulting edges is calculated by $|\cdot|$, and $\sigma(\cdot)$ is the sigmoid activation function. This produces smooth edge maps $\mathcal{E}_p$ and $\mathcal{E}_c$ for the person's body and clothing, respectively. Next, we compute the final interaction edges $\mathcal{E}_{int}$, which define the boundary regions where the fabric deformation is most prominent. These edges are determined by the spatial consistency between the person's body and clothing edges, incorporating both edge magnitude and orientation similarity:

$$\mathcal{E}_{int} = \mathbb{I}\left[(\mathcal{E}_p * \mathcal{G}_\sigma) \odot (\mathcal{E}_c * \mathcal{G}_\sigma) \odot \left(0.5 + 0.5\cos\theta_{pc}\right) > \tau\right], \quad (7)$$

where $*$ represents the convolution operation, applied here with a Gaussian kernel $\mathcal{G}_\sigma$, which is used to smooth the edge maps $\mathcal{E}_p$ and $\mathcal{E}_c$. $\cos\theta_{pc}$ measures the similarity in orientation between the person's body and clothing edges, quantifying how well aligned the edges are in terms of their geometric direction. τ is a threshold that determines the significance of the interaction edges.

(2) **Physics-based Attraction Field.** For each pixel $p = (x, y)$ in the image, we establish an **inverse-square attraction field** radiating toward the interior of the clothing, where the force at position p is determined by the sum of the contributions from all edge pixels $p_i \in \mathcal{E}_{\text{int}}$ (the interior edges). The force is directed towards each edge pixel, and its magnitude decays following the inverse-square law as the distance between p and p_i increases. The attraction field at position p is given by:

$$A(p) = \sum_{p_i \in \mathcal{E}_{int}} \frac{\alpha}{1 + \beta\|p - p_i\|_2^2} \cdot \frac{p_i - p}{\|p_i - p\|}, \quad (8)$$

where $\|p - p_i\|_2$ represents the Euclidean distance between the current pixel p and an edge pixel p_i, α controls the intensity of the attraction force, and β determines how rapidly the attraction force attenuates with distance.

The total displacement at a position p is accumulated by summing the displacement contributions from all edge pixels p_i. The displacement vector at p is given by:

$$\Delta(p) = \sum_{i=1}^{N} \frac{\alpha(p_i - p)}{1 + \beta\|p - p_i\|_2^2}. \quad (9)$$

This displacement accumulates the contributions of each edge pixel, scaled by a distance-dependent factor. It represents the total displacement at a point p due to the edge pixel p_i. To avoid the computational burden during training, the attraction field can be precalculated and loaded as labels.

(3) **Adaptive Texture Warping.** The final adversarial texture $\hat{\mathcal{T}}$ is generated through deformable grid sampling from the rotated and scaled texture $\mathcal{T}' = \mathcal{S}(\mathcal{T}, \text{bbox}(I))$. The warping mechanism adapts the sampling coordinates based on a learned displacement field, producing a spatially coherent deformation. Specifically, the warped texture at pixel position $p = (x, y)$ is given by:

$$\hat{\mathcal{T}}(x, y) = \sum_{i,j} \mathcal{T}'(i, j) \cdot \delta\left(\begin{bmatrix} x \\ y \end{bmatrix} + \eta\Delta(p) - \begin{bmatrix} i \\ j \end{bmatrix}\right), \tag{10}$$

where $\delta(\cdot)$ denotes the bilinear interpolation kernel centered at integer coordinates (i, j), $\Delta(p) \in \mathbb{R}^2$ is the displacement vector at location p, and $\eta \in \mathbb{R}$ is a scalar hyperparameter that controls the overall strength of the deformation applied to the sampling grid.

3.4 Blur Simulation

To comprehensively model the process of blur generation, we incorporate both local motion blur and global blur effects—including lens blur and Gaussian blur—into the attack pipeline. Considering the presence of image restoration modules in modern intelligent cameras, we further integrate a deblurring model into our pipeline. This allows the adversarial texture to simultaneously target the restoration process, thereby achieving a more thorough disruption of the overall system and potentially impairing downstream tasks.

We simulate motion blur as a convolution between the rendered image $\hat{I}$ and a parametric blur kernel B_θ:

$$\tilde{I} = (\hat{I} * B_\theta) \odot M_{Person} + \hat{I} \odot (1 - M_{Person}), \tag{11}$$

where M^{person} is the mask indicating the person region affected by blur. The kernel parameter $\theta = (k, \alpha, d)$ comprises three components: the kernel size k, which controls the extent of the blur; the motion direction angle α; and the directional bias coefficient d, which determines the motion polarity—positive for forward and negative for backward motion. These parameters are randomly sampled from predefined ranges to simulate diverse motion scenarios. The angular motion is converted into a 2D motion vector $\mathbf{v}$ as:

$$\mathbf{v} = \begin{cases} [\cos\alpha, \sin\alpha]^T, & d \geq 0 \\ [\cos(\alpha + \pi), \sin(\alpha + \pi)]^T, & d < 0 \end{cases} \tag{12}$$

The blur kernel $B_\theta(i, j)$ is constructed using a linear attenuation model along the direction $\mathbf{v}$:

$$B_\theta(i, j) = \frac{1}{\eta} \max\left(1 - \frac{|\mathbf{v} \cdot (i - c, j - c)|}{\|\mathbf{v}\|}, 0\right), \tag{13}$$

where $c = \frac{k-1}{2}$ is the kernel center, and η ensures $\sum_{i,j} B_\theta(i, j) = 1$.

This motion blur kernel acts as a directional low-pass filter, averaging pixel intensities along $\mathbf{v}$. Larger kernel sizes (k) induce more aggressive blurring, primarily attenuating high-frequency content such as texture edges and fine details.

3.5 High Frequency Loss

To ensure the robustness of adversarial patterns against motion blur and deblurring operations, we propose the *High Frequency Loss* ($\mathcal{L}_{HF}$), which suppresses high-frequency components while reinforcing low-frequency features. This design is motivated by the fact that high-frequency patterns are more susceptible to degradation from blurring and deblurring processes, whereas low-frequency components exhibit greater stability under such transformations. A theoretical justification for the preservation of low-frequency components during motion blur is provided in *Appendix Section A*. By imposing a constraint in the frequency domain, we control the expression of adversarial information exclusively within the low-frequency components, enhancing resistance to motion blur and deblurring models. The calculation begins by transforming the patch p into the frequency domain using a 2D Fast Fourier Transform (FFT), denoted as $\mathcal{F}$:

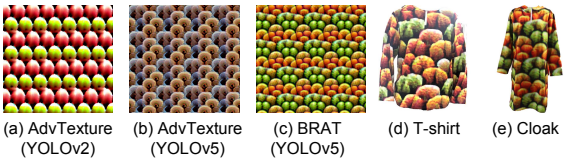


Figure 4: Visualization of different textures and cloths.

$$\Phi = \mathcal{F}(P), \tag{14}$$

where $\Phi \in \mathbb{C}^{C \times H \times W}$ denotes the complex frequency spectrum. To isolate the high-frequency components, we define:

$$\Phi_{high\_freq}(u, v) = \begin{cases} \Phi(u, v), & \text{if } d(u, v) \geq \rho \cdot d_{max} \\ 0, & \text{otherwise} \end{cases} \tag{15}$$

where $d(u, v)$ is the radial distance from the DC component, and d_{max} is the maximum frequency radius. Similarly, Φ_{low_freq} can be obtained by taking the complement of the high-frequency region.

The frequency components are then transformed back into the spatial domain using inverse FFT, denoted as $\mathcal{F}^{-1}$:

$$P_{high\_freq} = \mathcal{F}^{-1}(\Phi_{high\_freq}), \quad P_{low\_freq} = \mathcal{F}^{-1}(\Phi_{low\_freq}). \tag{16}$$

The high-frequency loss is defined as:

$$\mathcal{L}_{HF} = \frac{1}{CHW} \sum_{c,i,j} |P_{high\_freq}(c, i, j)|. \tag{17}$$

This term suppresses high-frequency components susceptible to motion blur, while implicitly enhancing more robust low-frequency features. By leveraging their complementarity, $\mathcal{L}_{HF}$ achieves both suppression and enhancement in a unified framework.

4 EXPERIMENT

4.1 Experiment Settings

4.1.1 Datasets. **1) Digital-space.** We employ the COCO dataset [20], from which we selectively filter images featuring individuals with more than 10 visible keypoints, resulting in a subset referred to as COCOPerson. We randomly select 5000 images for model training, while the test set consists of 754 images. **2) Physical-space.** We recruit three actors (height range: 168-188 cm) to collect physical test data. We captured 8,000 images in total across 10 different scenes (6 indoors and 4 outdoors). See *Appendix Section B* for details.

4.1.2 Baseline Methods. We perform a comprehensive comparison with all existing physical adversarial attack methods, prioritizing optimal attack performance over naturalness. These attack methods can be broadly categorized into two groups: 1) patch-based methods, including AdvPatch [28], AdvT-shirt [37] , AdvCloak [36], and T-SEA [15]; and 2) texture-based methods, such as AdvTexture [14].

Table 1: Digital-space quantitative comparison of different methods.

Mode	Method	Natural Expansion	*w/* Motion Blur		*w/o* Blur		*w/* Lens Blur		*w/* Both Blurs		Average	
			AP ↓	ASR ↑	AP ↓	ASR ↑	AP ↓	ASR ↑	AP ↓	ASR ↑	AP ↓	ASR ↑
Patch	Random Noise	–	92.2	13.9	92.4	13.0	92.2	11.5	91.8	14.0	92.2	13.1
	AdvPatch [28]	–	69.4	42.0	49.2	61.8	55.4	58.0	63.7	49.5	59.4	52.8
	AdvT-shirt [37]	–	82.9	21.2	77.1	24.4	82.0	19.1	83.6	20.3	81.4	21.3
	AdvCloak [36]	–	80.1	27.8	74.0	30.4	75.7	33.7	79.7	30.0	77.4	30.5
	T-SEA [15]	–	44.9	45.7	26.4	51.4	24.8	61.0	33.6	53.6	32.4	52.9
Texture	Random Noise	✓	92.3	13.8	92.3	13.1	92.3	12.6	92.4	11.7	92.3	12.8
	AdvPatch	✗	56.5	54.0	38.3	65.5	34.7	70.0	50.5	54.8	45.0	61.1
	AdvT-shirt	✗	75.2	29.8	61.8	38.8	67.8	36.2	74.1	32.6	69.7	34.4
	AdvCloak	✗	61.3	46.5	42.4	57.7	44.1	59.3	56.5	50.8	51.1	53.6
	T-SEA	✗	41.6	61.5	26.1	69.8	26.4	72.6	35.1	66.2	32.3	67.5
	AdvTexture [14]	✓	58.8	48.1	36.1	63.1	38.4	63.9	50.2	52.2	46.1	56.6
	FAB-Attack	✓	**31.0**	**71.5**	**25.2**	**74.9**	**24.0**	**74.5**	**28.0**	**73.3**	**27.1**	**73.6**

For patch-based attacks, we tile their patches to form textures with repeated patterns [14] for texture mode evaluations.

4.1.3 Victim Models. Our work encompasses both attacks on person detection models and image restoration models. **1)** For person detection, we use YOLOv5 [31] as a white-box model and select YOLOv3 [8], YOLOv8 [17], YOLO11 [18], Faster R-CNN [23], Mask R-CNN [10], CenterNet [43], and DETR [5] for black-box models. To ensure reproducibility and facilitate fair comparison, we employ the official pre-trained models (see *Appendix Section E*). **2)** For image restoration, focusing on image deblurring capabilities, we use NAFNet [7] as a white-box model and select Restormer [40], DeepRTF [21] and UFormer [34] as black-box models.

4.1.4 Metrics. Following [15, 28, 35], we utilize the Average Precision (AP%), a metric assessing the accuracy of the detection model, to evaluate the performance of attack methods. Lower AP values indicate better attack performance. The IoU threshold for AP during evaluation is set to 0.5. Another metric we employ is the Attack Success Rate (ASR%), calculated as $1 - TP'/TP$, where TP represents the number of True Positive samples without attacks, and TP' represents the number of True Positive samples with attacks. A higher ASR value signifies superior attack performance.

4.1.5 Implementation Details. During testing in digital space, we set the confidence score threshold as **0.001** for all detectors involved and set Non-Maximum Suppression (NMS) threshold as **0.4** for detectors who need it. During testing in physical space, we set the confidence score threshold at 0.5 and set no IoU threshold, which means that any detection boxes appearing in the human body will lead to a *failure of attack*. See *Appendix Section C* for more details.

4.2 Digital-Space Evaluation

In our evaluation, we consider two main perturbation application modes: patch mode and texture mode. In patch mode, adversarial patches follow common practice and are placed at the center of the target bounding box. In contrast, the texture mode employs our proposed Fabric-aware Texture Appliance (FTA) without Physics-inspired TPS (for fair comparison), wherein baseline patches are tiled into textures following the approach of [14]. If there is no additional clarification, the texture mode will be a default mode.

4.2.1 Comparison with Baselines. We first evaluate the attack performance of the proposed FAB-Attack under a white-box setting within digital space. To comprehensively illustrate how different blur-induced image degradations affect physical adversarial attacks, we conduct tests under four distinct scenarios: no blur, motion blur, lens blur, and a combination of both motion and lens blur. Here, random Gaussian blur is utilized to simulate lens blur effects.

Tab. 1 summarizes the performance of the baseline methods compared to our FAB-Attack method. It is noteworthy that motion blur significantly deteriorates the attack effectiveness of all baseline methods, regardless of whether they are applied in patch or texture mode. In contrast, our proposed FAB-Attack consistently maintains stable and superior performance, highlighting the critical impact of motion blur on physical adversarial attack workflows.

Natural Expansion in Tab. 1 denotes whether the corresponding perturbation can be naturally tiled into a fluent texture without pattern gaps. FAB-Attack and AdvTexture [14] both achieve a naturally expandable adversarial pattern, addressing the segment-missing problem, as visualization results shown in Fig. 4.

Compared to the substantial defensive effect of motion blur, lens blur introduces only minor performance degradation for all tested methods. Interestingly, lens blur appears to mitigate the adverse impact of motion blur when both blur types are combined. This may be attributed to the smoothing effect of lens blur, which reduces edge sharpness and directional features. Given that lens blur poses minimal threat to existing physical adversarial attack methods, we primarily focus on motion blur scenarios in subsequent evaluations.

4.2.2 Transferability of FAB-Attack in Digital Space. To evaluate the transferability of FAB-Attack, we conduct experiments across eight widely-used object detectors beyond YOLOv5, including both one-stage and two-stage detectors. The results are summarized in Tab. 2. We compare our approach against AdvTexture [14] and evaluate the attack performance under two conditions: *Clean* (without blur) and random *Motion Blur*, simulating real-world deployment

Table 2: Black-box evaluations in digital space.

Detector	Motion Blur				Clean			
	AdvTexture		FAB-Attack		AdvTexture		FAB-Attack	
	AP↓	ASR↑	AP↓	ASR↑	AP↓	ASR↑	AP↓	ASR↑
YOLOv5	58.8	48.1	**31.0**	**71.5**	36.1	63.9	**25.2**	**74.9**
YOLOv3	59.0	47.7	**35.7**	**67.1**	46.2	58.4	**28.9**	**71.9**
YOLOv8	68.5	39.5	**47.4**	**58.4**	56.6	47.4	**38.5**	**64.0**
YOLO11	72.3	36.8	**48.2**	**59.8**	58.4	53.2	**39.7**	**62.1**
Faster R-CNN	62.5	47.7	**41.0**	**60.7**	54.5	54.8	**32.6**	**67.2**
Mask R-CNN	58.3	51.9	**36.6**	**66.5**	49.3	54.6	**29.0**	**72.8**
CenterNet	64.4	44.6	**51.1**	**52.4**	53.1	52.1	**44.6**	**59.5**
DETR	65.6	38.2	**53.2**	**50.1**	61.4	38.4	**52.3**	**52.8**

Table 3: Evaluation of attack performance against image deblurrers in digital and physical spaces.

(a) Digital space

Deblurrer	FAB-Attack		AdvTexture		Clean	
	PSNR ↓	SSIM ↓	PSNR ↓	SSIM ↓	PSNR ↓	SSIM ↓
None	25.1	0.783	24.9	0.776	25.3	0.801
NAFNet	**25.1**	**0.804**	26.9	0.873	27.1	0.894
Restormer	**25.8**	**0.808**	27.3	0.856	28.9	0.887
DeepRFT	**24.9**	**0.790**	25.9	0.821	26.9	0.876
UFormer	**25.6**	**0.805**	26.5	0.844	27.8	0.877

(b) Physical space

Deblurrer	FAB-Attack		AdvTexture		Clean	
	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑
NAFNet	**36.1**	**0.991**	34.6	0.977	33.8	0.941
Restormer	**35.9**	**0.990**	34.7	0.980	34.4	0.980
DeepRFT	**36.0**	**0.987**	34.1	0.968	34.2	0.972
UFormer	**35.5**	**0.985**	35.0	0.979	34.4	0.971

scenarios. The results show that FAB-Attack consistently outperforms AdvTexture across all detectors in terms of both AP and ASR, demonstrating its transferability to various black-box models.

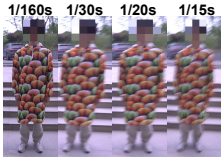
4.2.3 Against Image Restoration Model. FAB-Attack is specifically designed to target image restoration models in order to disrupt the deblurring process. To evaluate the effectiveness of this attack on image deblurring, we conduct a comprehensive experiment comparing the proposed FAB-Attack and AdvTexture [14] against four prominent image restoration models[7, 21, 33, 40]. In the experiment, we randomly apply global motion blur to images with textures generated by FAB-Attack and AdvTexture, and include a contrast group with no adversarial texture, referred to as "Clean". The results are presented in Fig. 3a. We report the PSNR and SSIM [32] between the deblurred images and their corresponding original, motion-blurred images, where a lower value indicates superior attack performance. The row labeled "Deblurrer: None" represents results obtained without applying any deblurring process after the motion blur is introduced. The closer the subsequent rows are to this baseline, the more effective the attack. Experimental results demonstrate that the proposed FAB-Attack consistently exhibits strong attack performance across multiple deblurring models.

Table 4: Demonstration of the methods involved in our physical-space evaluation.

Method	Training Paradigm	TPS type	HF loss	Motion Blur Modeling
FAB-Attack	texture	p-TPS	✓	✓
FAB-Attack†	texture	random TPS	✓	✓
FAB-Attack‡	patch	random TPS	✗	✓
AdvTexture	patch	random TPS	✗	✗

Table 5: The ASRs of different attacks in physical space with motion blur under YOLOv5 [31] (frontal view only).

Method	1/160s	1/30s	1/20s	1/15s
AdvTexture	95.1	79.5	72.9	61.9
FAB-Attack	**99.7**	**99.5**	**97.5**	**94.4**
FAB-Attack†	98.2	96.8	95.7	90.1
FAB-Attack‡	92.7	87.1	85.5	80.1



4.3 Physical-Space Evaluation

For physical-space evaluations, we fabricate multiple garments with different adversarial textures, including AdvTexture [14], our proposed FAB-Attack and two baselines for ablation study. The training condition details of the four adversarial textures can be found in Tab. 4. For each adversarial textures, we fabricate a cloak and a T-shirt (see Fig. 4). Cloaks will be default options if there is no additional clarification.

4.3.1 Under Motion Blur. To evaluate the robustness and attack effectiveness of the proposed FAB-Attack, we collected a total of 4,800 images across 10 different scenarios, each subjected to four levels of motion blur. During data collection, a cameraman intentionally moved the camera with regular amplitude and frequency to induce motion blur. The degree of blur was controlled by varying the shutter speed—longer exposure times resulted in more pronounced motion blur. The results are summarized in Tab. 5. FAB-Attack demonstrated robust attack performance across all blur levels. FAB-Attack†, which also incorporates High Frequency Loss and motion blur modeling, exhibited comparable robustness. In contrast, FAB-Attack‡, which omits High Frequency Loss, showed reduced effectiveness under motion blur. AdvTexture experienced the most significant performance degradation in the presence of motion blur.

4.3.2 Under Physical Variation. We conduct comprehensive experiments to further evaluate the robustness of the proposed FAB-Attack under physical-space variations. Our tests are divided into three main parts: 1) the attack performance of FAB-Attack at different distances, where we compute the average attack success rate (ASR) at eight distinct viewing angles for each distance (see Fig. 5a); 2) the attack performance of FAB-Attack under different forms (T-shirt or cloak) and various viewing points (see Fig. 6b); and 3) the attack performance of FAB-Attack during a series of complex actions performed by the wearer, to assess the ASR under various real fabric deformations (see Fig. 6). The results indicate that FAB-Attack demonstrates consistent and impressive attack performance across various viewing points at distances between 2 and 6 meters. This robustness is largely due to our FTA, which significantly bridges the domain gap between training and physical testing. In contrast, AdvTexture shows a noticeable decline in

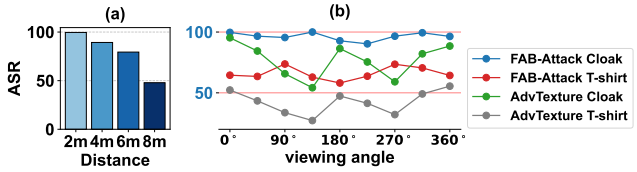


Figure 5: (a) The mASRs of different attacks at multiple angles with different distances. (b) The ASRs at multiple angles.

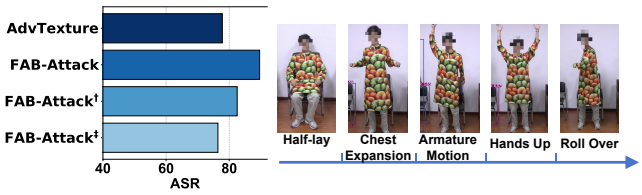


Figure 6: The mASRs under a sequence of actions.

Table 6: The mASRs in multiple angles across detectors.

Method	YOLOv3 [31]	YOLOv5 [8]	YOLOv8 [17]	YOLO11 [18]
AdvTexture	67.3	76.8	40.6	27.0
FAB-Attack	**92.1**	**96.4**	**90.3**	**82.9**

Table 7: Digital-space evaluation using FTA under YOLOv5.

Metric	AdvTexture	FAB-Attack	FAB-Attack†	FAB-Attack‡
AP ↓	41.3	**27.8**	32.0	42.8
ASR ↑	60.2	**74.5**	72.9	59.1

performance at certain viewing points. A similar trend is observed in the action sequence tests, where FAB-Attack outperforms in the presence of complex cloth warping caused by various actions. While FAB-Attack† exhibits slightly reduced stability under fabric variety, its use of cloth-region texturing during training still ensures strong overall performance. In comparison, FAB-Attack‡ shows inferior robustness under the same conditions, with performance comparable to that of AdvTexture.

4.3.3 Across Detectors. We further test the transferability of FAB-Attack comparing with AdvTexture [14], with the results presented in Tab. 6. The test involves 800 images captured across 10 different scenarios and 8 viewing points. We report the average ASR across all samples. The results demonstrate that FAB-Attack maintains stable attack performance across different detectors, significantly surpassing AdvTexture. This may be attributed to the use of model shake drop [15] during training.

4.3.4 Against Image Restoration Models. Finally, we investigate the effectiveness of FAB-Attack against image restoration models in physical space. We use the images from Sec. 4.3.1 as our test samples. Unlike evaluations in digital space, there is no ground truth in physical space. Therefore, we compute the PSNR and SSIM between the input and output of the deblurring models, where higher values indicate better attack performance. The results, presented in Tab. 3b, clearly demonstrate that FAB-Attack remains effective against deblurring models in physical space, regardless of whether the models are black-box or white-box.

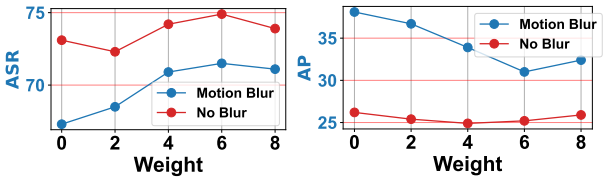


Figure 7: Ablation study on High Frequency Loss.

4.4 Ablation Study

4.4.1 Analysis of High Frequency Loss. To thoroughly evaluate the effectiveness of the proposed High Frequency Loss ($\mathcal{L}_{HF}$), we vary its loss weight during training to generate multiple variants of FAB-Attack. All variants are evaluated under the same settings described in Sec. 4.2.1, with the results illustrated in Fig. 7. As shown in the figure, increasing the HF loss weight significantly reduces the performance gap between the Motion Blur and No Blur settings in terms of both ASR and AP, demonstrating the role of $\mathcal{L}_{HF}$ in improving robustness to motion blur. However, when the weight reaches 8, the overall attack strength begins to degrade. Therefore, we adopt a weight of 6 as the default setting in our experiments. The physical-space results in Sec. 4.3.1 further corroborate the contribution of $\mathcal{L}_{HF}$ to motion blur robustness, where the method using $\mathcal{L}_{HF}$—FAB-Attack and FAB-Attack†—exhibit a significantly smaller performance drop under severe motion blur compared to FAB-Attack‡ without $\mathcal{L}_{HF}$.

4.4.2 Effect of FTA on Attack. To investigate the effectiveness of FTA, we evaluate different methods **using full FTA** in the digital space under the No Blur setting. As shown in Tab. 7, FAB-Attack, which integrates texture-based training and p-TPS, achieves the strongest attack performance. FAB-Attack†, which also employs texture training but replaces p-TPS with random TPS, exhibits a noticeable drop in attack effectiveness. FAB-Attack‡ and AdvTexture, which both rely solely on adversarial patches with random TPS, show a further decline in performance. Similar trends are observed in the physical-space analysis presented in Sec.4.3.2. The results highlight the importance of both the texture-based training paradigm and the use of p-TPS in enhancing attack effectiveness.

5 CONCLUSION

In this paper, we introduce FAB-Attack, an effective adversarial framework targeting person detectors that remains robust under real-world physical variations, such as clothing deformation and motion blur. Beyond attacking the detector directly, FAB-Attack also disrupts deblurring models, further impairing downstream perception. Unlike prior attacks leaving a great domain gap between training and physical testing, we propose FTA, a physics-grounded method that renders fabric-realistic textures on clothing in real-world datasets, achieving better alignment with physical-world dynamics. To enhance robustness under motion-induced corruption, FAB-Attack suppresses high-frequency texture components that are prone to degradation. Extensive experiments validate the effectiveness of FAB-Attack across both digital and physical settings, demonstrating superior robustness and transferability.

REFERENCES

[1] Anish Athalye, Logan Engstrom, Andrew Ilyas, and Kevin Kwok. 2018. Synthesizing robust adversarial examples. In *International conference on machine learning*. PMLR, 284–293.

[2] David Baraff and Andrew Witkin. 2023. *Large Steps in Cloth Simulation* (1 ed.). Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3596711.3596792

[3] Tom B Brown, Dandelion Mané, Aurko Roy, Martín Abadi, and Justin Gilmer. 2017. Adversarial patch. In *NeurIPS Workshop*.

[4] Zikui Cai, Xinxin Xie, Shasha Li, Mingjun Yin, Chengyu Song, Srikanth V Krishnamurthy, Amit K Roy-Chowdhury, and M Salman Asif. 2022. Context-aware transfer attacks for object detection. In *AAAI Conference on Artificial Intelligence*.

[5] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. 2020. End-to-End Object Detection with Transformers. In *European conference on computer vision*.

[6] Mengying Chang, Huizhi Xu, and Yuanming Zhang. 2025. Low light recognition of traffic police gestures based on lightweight extraction of skeleton features. *Neurocomputing* 617 (2025), 129042.

[7] Liangyu Chen, Xiaojie Chu, Xiangyu Zhang, and Jian Sun. 2022. Simple Baselines for Image Restoration. *arXiv preprint arXiv:2204.04676* (2022).

[8] Ali Farhadi and Joseph Redmon. 2018. Yolov3: An incremental improvement. In *Computer vision and pattern recognition*, Vol. 1804. Springer Berlin/Heidelberg, Germany, 1–6.

[9] Amira Guesmi, Ruitian Ding, Muhammad Abdullah Hanif, Ihsen Alouani, and Muhammad Shafique. 2024. Dap: A dynamic adversarial patch for evading person detectors. 24595–24604.

[10] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. 2017. Mask r-cnn. In *IEEE/CVF International Conference on Computer Vision*.

[11] Ngoc Quy Hoang, Seungbo Shim, Seonghun Kang, and Jong-Sub Lee. 2024. Anomaly detection via improvement of GPR image quality using ensemble restoration networks. *Automation in Construction* 165 (2024), 105552.

[12] Yu-Chih-Tuan Hu, Bo-Han Kung, Daniel Stanley Tan, Jun-Cheng Chen, Kai-Lung Hua, and Wen-Huang Cheng. 2021. Naturalistic Physical Adversarial Patch for Object Detectors.

[13] Zhanhao Hu, Wenda Chu, Xiaopei Zhu, Hui Zhang, Bo Zhang, and Xiaolin Hu. 2023. Physically realizable natural-looking clothing textures evade person detectors via 3d modeling. 16975–16984.

[14] Zhanhao Hu, Siyuan Huang, Xiaopei Zhu, Fuchun Sun, Bo Zhang, and Xiaolin Hu. 2022. Adversarial Texture for Fooling Person Detectors in the Physical World. 13307–13316.

[15] Hao Huang, Ziyan Chen, Huanran Chen, Yongtao Wang, and Kevin Zhang. 2023. T-sea: Transfer-based self-ensemble attack on object detection. 20514–20523.

[16] Lifeng Huang, Chengying Gao, Yuyin Zhou, Cihang Xie, Alan L Yuille, Changqing Zou, and Ning Liu. 2020. Universal physical camouflage attacks on object detectors. 720–729.

[17] Glenn Jocher, Ayush Chaurasia, and Jing Qiu. 2023. *Ultralytics YOLOv8*. https://github.com/ultralytics/ultralytics

[18] Glenn Jocher and Jing Qiu. 2024. *Ultralytics YOLO11*. https://github.com/ultralytics/ultralytics

[19] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. 2021. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*. 1833–1844.

[20] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*.

[21] Xintian Mao, Yiming Liu, Fengze Liu, Qingli Li, Wei Shen, and Yan Wang. 2023. Intriguing findings of frequency selection for image deblurring. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37. 1905–1913.

[22] Federico Nesti, Giulio Rossolini, Saasha Nair, Alessandro Biondi, and Giorgio Buttazzo. 2022. Evaluating the robustness of semantic segmentation for autonomous driving against real-world adversarial patch attacks. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2280–2289.

[23] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems* 28 (2015).

[24] Mahmood Sharif, Sruti Bhagavatula, Lujo Bauer, and Michael K Reiter. 2016. Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition. In *Proceedings of the 2016 acm sigsac conference on computer and communications security*. 1528–1540.

[25] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. 2014. Intriguing properties of neural networks. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, Yoshua Bengio and Yann LeCun (Eds.). http://arxiv.org/abs/1312.6199

[26] Jia Tan, Nan Ji, Haidong Xie, and Xueshuang Xiang. 2021. Legitimate Adversarial Patches: Evading Human Eyes and Detection Models in the Physical World. In *ACM International Conference on Multimedia*.

[27] Jia Tan, Nan Ji, Haidong Xie, and Xueshuang Xiang. 2021. Legitimate Adversarial Patches: Evading Human Eyes and Detection Models in the Physical World. In *Proceedings of the 29th ACM International Conference on Multimedia* (Virtual Event, China) *(MM '21)*. Association for Computing Machinery, New York, NY, USA, 5307–5315. https://doi.org/10.1145/3474085.3475653

[28] Simen Thys, Wiebe Van Ranst, and Toon Goedemé. 2019. Fooling automated surveillance cameras: adversarial patches to attack person detection. 0–0.

[29] Dai Quoc Tran, Armstrong Aboah, Yuntae Jeon, Maged Shoman, Minsoo Park, and Seunghee Park. 2024. Low-light image enhancement framework for improved object detection in fisheye lens datasets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7056–7065.

[30] Zhengzhong Tu, Hossein Talebi, Han Zhang, Feng Yang, Peyman Milanfar, Alan Bovik, and Yinxiao Li. 2022. Maxim: Multi-axis mlp for image processing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 5769–5780.

[31] Ultralytics. 2020. yolov5. https://github.com/ultralytics/yolov5. Accessed: 2024-03-04.

[32] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing* 13, 4 (2004), 600–612.

[33] Zhendong Wang, Xiaodong Cun, Jianmin Bao, Wengang Zhou, Jianzhuang Liu, and Houqiang Li. 2022. Uformer: A general u-shaped transformer for image restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 17683–17693.

[34] Zhendong Wang, Xiaodong Cun, Jianmin Bao, Wengang Zhou, Jianzhuang Liu, and Houqiang Li. 2022. Uformer: A General U-Shaped Transformer for Image Restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 17683–17693.

[35] Hui Wei, Hao Tang, Xuemei Jia, Zhixiang Wang, Hanxun Yu, Zhubo Li, Shin'ichi Satoh, Luc Van Gool, and Zheng Wang. 2024. Physical adversarial attack meets computer vision: A decade survey. (2024).

[36] Zuxuan Wu, Ser-Nam Lim, Larry S Davis, and Tom Goldstein. 2020. Making an invisibility cloak: Real world adversarial attacks on object detectors. Springer, 1–17.

[37] Kaidi Xu, Gaoyuan Zhang, Sijia Liu, Quanfu Fan, Mengshu Sun, Hongge Chen, Pin-Yu Chen, Yanzhi Wang, and Xue Lin. 2020. Adversarial t-shirt! evading person detectors in a physical world.

[38] Qiuling Xu, Guanhong Tao, Siyuan Cheng, and Xiangyu Zhang. 2021. Towards feature space adversarial attack by style perturbation. In *AAAI Conference on Artificial Intelligence*, Vol. 35. 10523–10531.

[39] Darren Yu Yang, Jay Xiong, Xincheng Li, Xu Yan, John Raiti, Yuntao Wang, HuaQiang Wu, and Zhenyu Zhong. 2018. Building Towards" Invisible Cloak": Robust Physical Adversarial Attack on YOLO Object Detector. In *2018 9th IEEE Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON)*. IEEE, 368–374.

[40] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. 2022. Restormer: Efficient Transformer for High-Resolution Image Restoration. In *CVPR*.

[41] Shibo Zhang, Yushi Cheng, Wenjun Zhu, Xiaoyu Ji, and Wenyuan Xu. 2023. CAPatch: Physical Adversarial Patch against Image Captioning Systems. (2023).

[42] Xu Zhao, Wenchao Ding, Yongqi An, Yinglong Du, Tao Yu, Min Li, Ming Tang, and Jinqiao Wang. 2023. Fast Segment Anything. arXiv:2306.12156 [cs.CV]

[43] Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl. 2019. Objects as Points. In *arXiv preprint arXiv:1904.07850*.